RESEARCH ARTICLE

Evaluation of population stratification adjustment using genome-wide or exonic variants

Yuning Chen[1] | Gina M. Peloso[1] | Ching-Ti Liu[1] |
Anita L. DeStefano[1,2] | Josée Dupuis[1]

[1]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts

[2]Department of Neurology, Boston University School of Medicine, Boston, Massachusetts

Correspondence
Yuning Chen, Department of Biostatistics, Boston University School of Public Health, Boston, MA.
Email: yuningch@bu.edu

Abstract

Population stratification may cause an inflated type-I error and spurious association when assessing the association between genetic variations with an outcome. Many genetic association studies are now using exonic variants, which captures only 1% of the genome, however, population stratification adjustments have not been evaluated in the context of exonic variants. We compare the performance of two established approaches: principal components analysis (PCA) and mixed-effects models and assess the utility of genome-wide (GW) and exonic variants, by simulation and using a data set from the Framingham Heart Study. Our results illustrate that although the PCs and genetic relationship matrices computed by GW and exonic markers are different, the type-I error rate of association tests for common variants with additive effect appear to be properly controlled in the presence of population stratification. In addition, by considering single nucleotide variants (SNVs) that have different levels of confounding by population stratification, we also compare the power across multiple association approaches to account for population stratification such as PC-based corrections and mixed-effects models. We find that while these two methods achieve a similar power for SNVs that have a low or medium level of confounding by population stratification, mixed-effects model can reach a higher power for SNVs highly confounded by population stratification.

KEYWORDS
GWAS, mixed-effects model, PCA, population stratification

1 | INTRODUCTION

Genome-wide association studies (GWAS) have been proven to be a useful tool to discover single nucleotide variants (SNVs) associated with complex traits (Bush & Moore, 2012; Spencer, Su, Donnelly, & Marchini, 2009; Visscher, Brown, McCarthy, & Yang, 2012; Visscher et al., 2017). Population stratification, which is the allele frequency difference between cases and controls due to ancestry difference, occurs when there are multiple population groups within a sample (Hirschhorn et al., 2004), which has become a common study design to discover novel genetic variations and achieve higher statistical power. The association test results can be affected by an inflated type-I error caused by population stratification. Current methods for correcting population stratification include principal components (PCs) of the genotypes (Patterson, Price, & Reich, 2006; Price et al., 2006; Tucker,

Price, & Berger, 2014), genomic control factor (Devlin & Roeder, 1999; Wang, 2009), linear mixed-effects (LME) and generalized linear mixed-effects (GLME) model using an empirical kinship matrix (Chen et al., 2016; Kang et al., 2008; Kang et al., 2010; Lippert et al., 2011; Loh et al., 2015; Yang, Lee, Goddard, & Visscher, 2011; Zhang et al., 2010; Zhou & Stephens, 2012), structured association (Pritchard, Stephens, & Donnelly, 2000), and PC-AiR (Conomos, Miller, & Thornton, 2015) and PC-Relate (Conomos, Reiner, Weir, & Thornton, 2016) on studies with family structure and cryptic relatedness.

PC correction has been widely used in GWAS. Population stratification can be corrected by including genetic PCs as covariates in a linear regression model for continuous traits or a logistic regression model for binary traits. In contrast, in the genomic control method, an overall inflation factor is used to adjust the association test statistic at every SNV. Some SNVs have a bigger difference in allele frequencies across different populations, while some SNVs are not affected by population stratification. The overall inflation factor treats all SNVs the same and hence it may over-adjust SNVs with small differentiation across ancestral populations and under-adjust SNVs with strong differentiation. Yet another approach, mixed-effects models, utilize a variance component method to model genetic relationships. The model includes an empirically estimated genetic relatedness matrix (GRM), which takes advantage of the high-density genotype information, and estimates the variance parameters under the null model assuming the effect of any given variant on the phenotype is very small. The SNV effect is modeled as a fixed-effect and a random effect is included to model the relatedness among study participants. The structured association method adjusts for population stratification by assigning samples to subpopulation clusters and combines the association results of each cluster. This approach is highly sensitive to the number of subpopulation clusters and has intensive computational cost for large data sets, such as GWAS. Lastly, PC-AiR provides an accurate and robust genetic ancestry inference in the presence of both related and unrelated individuals. A subset of unrelated individuals that is representative of all ancestries in the sample is first identified through a computationally efficient algorithm. Then a PCA is performed on the genotype data from this subset of unrelated ancestry representative individuals, and PCs for all remaining individuals are computed based on their genetic similarities. PC-Relate provides accurate estimates of genetic relatedness by utilizing PCs calculated from PC-AiR to separate genetic correlations due to the sharing of recent ancestors from the genetic correlations due to more distant common ancestry.

These population stratification adjustment approaches were developed in the context of GWAS, which include common variants across the whole genome. However, the performance of these methods in association analyses has not been evaluated in studies with exonic variants. Whether exonic genotypes are sufficient to appropriately model population stratification is a question in studies without genome-wide (GW) genotypes, when PCs and GRM can only be computed using exonic variants.

Some studies have compared using exonic and GW variants under various contexts. Belkadi et al. (2016) found a strong correlation between PCs computed using GW and Whole Exome Sequencing (WES) variants, and that an accurate estimation of population stratification can be obtained using high-quality WES variants with minor allele frequency (MAF) > 2%. However, they did not compare type-I error and power in association tests with PC adjustment. Gazal et al. (2016) evaluated the performance of linkage analysis using GW and WES variants and showed that they had similar performance in excluding genomic regions with false-positive candidate causal variants. Smith et al. (2011) also demonstrated that accurate genetic linkage mapping can be performed using WES data. Kancheva et al. (2016) showed that WES variants can provide high specificity and sensitivity for the detection of homozygous regions in consanguineous families when using GW variants as reference. Eu-Ahsunthornwattana et al. (2014) compared kinship matrices computed using different numbers of SNVs and different software in 348 Brazilian families, and found that the kinship coefficients computed using ~50,000 SNVs were highly similar to those computed using ~545,000 SNVs.

In this paper, we focus on the PC-based and mixed-effects-model-based population stratification adjustment methods as genomic control and structured association have been shown not to be effective (Holland et al., 2006; Kang et al., 2008; Zhao et al., 2007). There are two concerns with models using exome-computed PCs and GRM: (a) The PCs and GRM can only capture the genetic information in exons and off-target variants near exons, while PCs and GRM computed using GW variants are able to capture the genetic information contained on the whole genome, hence they should be more accurate than exome-computed PCs and GRM; and (b) because the number of variants in exons is usually smaller than that in whole genome, exome-computed PCs and GRM may contain less information than GW-computed PCs and GRM due to the inclusion of fewer markers in the computation. The potential loss of using exonic variants can lead to insufficient adjustment for population stratification.

Our goals are to compare PCs and GRM computed using GW and exonic variants, to examine the effect of

the two sets of PCs and GRM on association analysis results of common SNVs with additive effect, and to evaluate performance of PC-based and LME/GLME models. We use simulations to compare GW- or exome-computed PCs and GRM in terms of genomic control factor, type-I error rate and power in association analyses. A real data application using data from the Framingham Heart Study (FHS) is used to compare the association results between height and rs2322659, an association known to be caused by population stratification in the *LCT* gene (Groop et al., 2005).

2 | METHODS

2.1 | Genotypes

We use the genotypes from the 1000 Genome Project Phase 3 (Auton et al., 2015) data set. GW variants are selected based on Illumina HumanHap300K BeadChip which is designed using the International HapMap Project (International HapMap Consortium, 2003) data of individuals from Northern and Western Europeans from Utah (CEU; Collins, 2007). Exonic variants are annotated using the EPACTS Version 3.2.6 (http://genome.sph.umich.edu/wiki/EPACTS) annotation function based on GENCODE V7 transcripts. We first apply the following quality control (QC) filters on all selected GW and exonic SNVs in the 1000G data set (combining all ancestries): MAF $\geq$ 1%, call rate $\geq$ 99%, Hardy–Weinberg equilibrium *p*-value > .0001. Variants passing these QC filters are used for evaluation of the type-I error rate in association tests. For PCs and GRM computation, we then select a subset of these SNVs based on the additional QC criterion: MAF $\geq$ 5% and only one SNV of each pair of SNVs with LD $r^2 > .5$ in a 50 SNVs window.

2.2 | PCs and GRM computation

PCs and two types of GRM, identity-by-state (IBS) and Balding–Nichols (BN) kinship matrix (Eu-Ahsunthornwattana et al., 2014) are computed using three sets of variants: (a) GW variants; (b) exonic variants; and (c) a randomly selected subset of GW variants that has the same number of variants as the exonic set. The inclusion of the third set of variants above is to eliminate the difference in the number of variants used in PCs and GRM calculation. IBS kinship matrix measures the proportion of alleles IBS between each pair of participants. In BN kinship matrix, the genetic relationship between individuals i and j is estimated through

$\sigma_{ij} = \frac{1}{2} \cdot \frac{1}{m} \sum_{k=1}^{k=m} \frac{(x_{k,i} - 2\hat{p}_k)(x_{k,j} - 2\hat{p}_k)}{2\hat{p}_k(1-\hat{p}_k)}$, where m is the total number of variants, $x_{k,i}$ and $x_{k,j}$ are the genotypes of individuals i and j and $\hat{p}_k$ is the MAF of the kth variant. This is equivalent to multiplying by ½ the GRM computed in GCTA (Yang et al., 2011). All PCs and GRM are computed using EIGENSTRAT (Price et al., 2006) and EMMAX (Kang et al., 2010), respectively.

2.3 | Simulation of study participants

We select founders from the 1000G Phase 3 data set to generate simulated data from two ethnic groups: European ancestry (EA) and African ancestry (AA). EA founders are from five cohorts: Finnish in Finland (FIN), Utah Residents with Northern and Western European Ancestry (CEU), British in England and Scotland (GBR), Iberian Population in Spain (IBS), and Toscani in Italy (TSI). AA founders are from seven cohorts: Americans of African Ancestry in Southwest USA (ASW), African Caribbean in Barbados (ACB), Luhya in Webuye, Kenya (LWK), Esan in Nigeria (ESN), Gambian in Western Divisions in the Gambia (GWD), Mende in Sierra Leone (MSL), and Yoruba in Ibadan, Nigeria (YRI). We then separately compute PCs on two sets of study participants: (a) using both EA and AA founders (EA + AA); and 2) restricting our analysis to EA founders only (EA only) to generate a data set with more subtle ancestry differences between subpopulations. Based on the clustering pattern (Figure 1), we divide the combined EA + AA 1000G founders into six groups (Group 1 = FIN; Group 2 = CEU + GBR; Group 3 = IBS + TSI; Group 4 = ASW + ACB; Group 5 = LWK; and Group 6 = ESN + GWD + MSL + YRI), while we do not combine cohorts in the analysis restricted to EA only, and simply refer to FIN, CEU, GBR, IBS, and TSI as groups 1–5, respectively. Next, within each group, we generate genotypes for 333 simulated individuals for the combined EA + AA analysis for each group and 400 individuals in each group for EA only analysis while maintaining the same linkage disequilibrium (LD) pattern as observed in each population between variants using the software Hapgen2 (Su, Marchini, & Donnelly, 2011). While one chromosome is simulated at a time, the software utilizes the Li-Stephens model (Li & Stephens, 2003) which relates the distribution of sampled haplotypes to the recombination rate. Through a Hidden Markov Model, each simulated haplotype is generated as an imperfect mosaic of the known haplotypes in the 1000 Genome and haplotypes that have already been simulated. To reconstitute the simulated haplotypes using mosaic fragments, the probability of change in the mosaic pattern is computed using the recombination rate estimates (Spencer
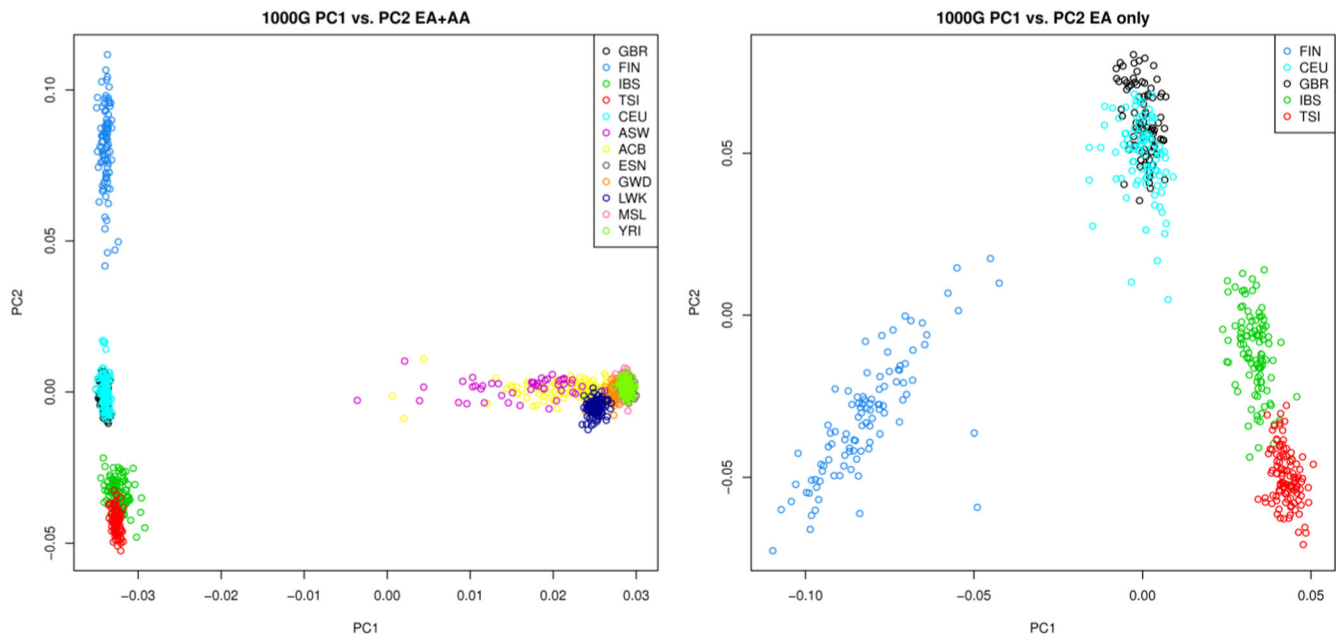
FIGURE 1 Population structure in 1000G Phase 3 data. The grouping in EA + AA samples is: Group 1 = FIN; Group 2 = CEU + GBR; Group 3 = IBS + TSI; Group 4 = ASW + ACB; group 5 = LWK; Group 6 = ESN + GWD + MSL + YRI. The grouping in EA only samples is: Group 1 = FIN; Group 2 = CEU; Group 3 = GBR; Group 4 = IBS; Group 5 = TSI. AA, African ancestry; ACB, African Caribbean in Barbados; ASW, Americans of African Ancestry in Southwest USA; CEU, Utah Residents with Northern and Western European Ancestry; EA, European ancestry; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GWD, Gambian in Western Divisions in the Gambia; IBS, Iberian Population in Spain; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; TSI, Toscani in Italy; YRI, Yoruba in Ibadan, Nigeria

et al., 2009). In total, 1,998 (six groups of 333) and 2,000 (five groups of 400) individuals are generated for the EA + AA, and EA only analyses, respectively.

2.4 | Comparison of PCs and GRM generated using GW and exonic variants

To compare PCs, we first consider a plane developed by the first two PCs because in the analysis with EA and AA samples, the top two PCs explain 7.52% and 0.31% of the variance, with corresponding eigenvalues equal to 84.37 and 3.50. Although these eigenvalues drop to 4 and 1.8 when considering EA samples only and they explain 0.8% and 0.35% of the total variance, the top two PCs still explain far more variance than other PCs. A centroid is defined by the mean vector of the first two PCs for each subgroup. Then the Euclidean distance from the group-specific centroid to each simulated sample is computed and the Wilcoxon signed-rank test is performed to compare the distance difference computed using GW variants, exonic variants, and the subset of GW variants which has the same number of variants as the exonic set. To compare GRMs, we compute the Pearson correlation between the kinship coefficients contained in the GRM.

2.5 | Comparison of genomic control factor and Type-I error rate

To examine genomic control factor, λ_{GC}, defined as the ratio of the median observed test statistic to the expected test statistic under the null hypothesis, and Type-I error rate, the continuous phenotype values are assigned based on a prespecified group-specific mean level. A random error term which follows a normal distribution is then added to the assigned mean levels. In the test of a binary outcome, a total of 999 or 1,000 randomly selected cases are generated for EA and AA, and EA only data, respectively. The prevalence of the simulated disease are set to 2%, 5%, 8%, 10%, 15%, and 18% for the six EA and AA groups, while they are 2%, 5%, 8%, 10%, and 15% for the five EA groups. The continuous and binary phenotypes are not associated with any simulated SNVs. We perform association analyses on all GW and exonic variants with MAF ≥ 1% regardless of whether or not they are included in the PC calculation to mimic what is typically done in GWAS. Variants with MAF < 1% are not included due to low statistical power to detect an association with low-frequency SNVs as well as inflated significance for binary traits with unbalanced designs (Ma, Blackwell, Boehnke, Scott, & GoT2D investigators, 2013). PCs and

GRM computed using GW and exonic SNVs are used for population stratification adjustment.

Four linear/logistic regression models are performed: (a) $Y \sim SNV$, an unadjusted model; (b) $Y \sim SNV + Group$, a model adjusted for the true group assignment, which is used as the gold standard; (c) $Y \sim SNV + PCs_{GW}$, a model adjusted for first 10 PCs (Price et al., 2006; Reed et al., 2015) computed from GW variants; and (d) $Y \sim SNV + PCs_{Exome}$, a model adjusted for the first 10 PCs computed from exonic variants. Besides these four fixed-effect models, four LME/GLME models, $Y \sim SNV + Kin_{GW/Exome, IBS/BN}$, with IBS/BN kinship matrix computed using GW/exonic variants are also performed. In addition, we also include PCs in the GLME models, $Y \sim SNV + PCs_{GW/Exome} + Kin_{GW/Exome, IBS/BN}$, for binary outcome to evaluate the performance of using both PCs and GRM to adjust for population stratification. The association analyses are repeated 500 times and performed using PLINK (Purcell et al., 2007) for PC-adjusted models, EMMAX and the R package GENESIS (Conomos et al., 2019) for mixed-effects models of continuous and binary traits, respectively.

2.6 | Power evaluation

To evaluate power, we select two sets of 30 SNVs (Table A1 in Appendix) to be associated with the phenotype for EA + AA and EA only analysis, respectively. The SNVs are selected based on their level of confounding from population stratification, as indicated by their PC weights for PC1 (EA + AA), or PC1 and PC2 (EA only). We consider PC1 alone in the analysis with EA + AA samples because PC1 explains far more variance than the rest of the PCs, as described above, whereas both PC1 and PC2 are considered as the population stratification is more subtle in the analysis restricted to EA samples only. To select the 30 SNVs, we first rank the absolute value of the weight of the first PC (EA + AA) or the weights of the first two PCs (EA only) in GW and exonic sets separately for each SNV present in both GW and exonic set. Then the ranks are added up and the top five SNVs, the bottom five SNVs, and five SNVs at each of the four percentiles: 25th, 50th, 75th, and 90th of the added rank are selected to have an effect on the simulated phenotypes. We refer the top five SNVs (higher rank and hence lower PC weights) as SNVs at 0th percentile of the added rank to indicate that they are not confounded by population stratification, and refer the bottom five SNVs (lower rank and hence higher PC weights) as SNVs at 100th percentile of the added rank to indicate that they are highly confounded.

By setting the proportion of the phenotypic variance explained by a SNV to $R^2 = 0.5\%$, the effect size β is set to be equal in all studies and it is computed using $\beta = \sqrt{\frac{R^2}{2p(1-p)}}$, where p is the MAF obtained across all groups. A normally distributed random error term is added to the linear combination of $\beta's$ and genotypes, with a group-specific mean and variance equal to 1. The associated binary trait is then generated by assigning samples whose continuous trait value is above the 90th percentile of all samples as cases to achieve a 10% population prevalence. We evaluate the same analysis models used for the Type-I error rate evaluation at α level of 1×10^{-3}.

2.7 | Comparison using FHS GW and exome chip data

In 1948, FHS enrolled its first participants, the Original Cohort with 5,209 individuals, from the town of Framingham, MA. These participants were aged between 30 and 62 at baseline and underwent a detailed physical examination, lifestyle interviews, and laboratory tests every 2 years to discover the risk factors of cardiovascular disease. The Offspring cohort comprised of 5,124 participants was recruited in 1971. These participants consist of the children and spouses of the children of the Original Cohort participants and attended physical exams approximately every 4 years. The third generation (Gen3) Cohort, which consists of the grandchildren of the Original Cohort and children of the Offspring cohort, was enrolled in 2002.

We perform a comparison using FHS GW and exome chip data with height as the outcome. Height was collected at Exam 1 for all individuals in the three cohorts. FHS participants in the SNP Health Association Resource (SHARe) were genotyped using the Affymetrix 500 K + 50 K MIPS chip. As part of the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium, exome chip (EC) variants were genotyped with the Illumina HumanExome BeadChip (Grove et al., 2013). The same filtering criteria are used to select variants for PC and GRM computation: $MAF \geq 5\%$, call rate $\geq 99\%$, Hardy–Weinberg equilibrium p-value > .0001 and only one SNV of each pair of SNVs with LD $r^2 > .5$ in a 50 SNVs window. Unrelated individuals are selected based on known family structures to compute the PC weights of each SNV, then PCs are projected on related participants (Price, Zaitlen, Patterson, & Reich, 2010). We test the association between 299,610 SNVs with $MAF \geq 5\%$ in the SHARe genome-wide variants set and the residual of

height computed with adjustment for sex, age, age^2, and a cohort indicator to account for the difference in mean phenotype across cohorts. We also report the result for SNV rs2322659, which is known for its spurious association with height due to population stratification. Two linear regression models, $Y \sim SNV + PCs_{GW}$ and $Y \sim SNV + PCs_{EC}$, and four mixed-effects models using GW-computed or EC-computed IBS and BN kinship matrix, respectively, are performed with adjustment for the first 10 PCs. In addition, we also apply PC-AiR, a robust population stratification inference in the presence of known or cryptic relatedness, and PC-Relate, measures of recent genetic relatedness in the presence of unspecified structure, to the GW and EC data sets. We use the KING-robust estimator in the PC-AiR calculation and then provide the PC-AiR results to the PC-Relate algorithm to get a GRM adjusted for population stratification. We include two mixed-effects models using GW- or EC-computed PC-AiR and PC-Relate GRM. We then compare the effect size estimate and p-value for each model.

3 | RESULTS

We select 498 unrelated EA individuals from five populations and 657 unrelated AA individuals from seven populations in 1000G Phase 3 data set (Table 1) as the reference sample in our simulations. We generate a total sample size of ~2,000 subjects in each simulation replicate. A total of 439,601 SNVs with MAF ≥ 1% for EA + AA, and 516,250 for EA only, pass the QC filters for Type-I error evaluation. Among 439,601 SNVs in the EA + AA analysis, 180,472 SNVs in the GW set and 66,166 SNVs in the exonic set pass the additional filtering criteria for including in the computation of PCs and GRM calculation. A total of 172,330 SNVs in the GW set and 57,584 SNVs in the WES set are included to compute PCs and GRM in the EA only analysis.

We make a pairwise comparison on distance difference computed using GW variants, exonic variants (referred to as Exome below) and a randomly selected subset of GW variants which has the same number of variants as the exonic set (referred to as Random below) using the Wilcoxon signed-rank test on the Euclidean distance from the group-specific centroid for each simulation iteration. Among 500 iterations, p-values of the Wilcoxon signed-rank test when comparing PCs computed from GW versus exonic variants, and GW versus Random are highly significant: The largest p-values (least significant) are 2.9×10^{-12} and 1.6×10^{-17} in the EA + AA data set, and 2.1×10^{-4} and 0.01 in the EA only data set, for the difference between GW and Exome, and GW and Random, respectively, which indicates that GW-computed PCs are

significantly different from both exome-computed PCs and PCs computed using Random. However, when restricting the GW set to have the same number of SNVs as the exonic set (comparing Random vs. Exome), the p-values are much less significant: only 100 iterations for EA + AA analysis and 26 for EA only analysis have p-value < .05, and 35 iterations for EA + AA analysis and five for EA only analysis have p-value < .01. These results show that the differences between GW-computed and exome-computed PCs are mainly due to the number of variants included in the calculation.

In the comparison of kinship coefficients (Figure 2), Pearson correlations demonstrate that kinship measures computed using GW and exonic variants are highly correlated when there are only EA samples, as shown in the right panel. In the presence of EA and AA samples, the kinship coefficients form three clusters corresponding to EA-EA, EA-AA, and AA-AA pairs. The correlation between Exome and GW, and Exome and the randomly selected set of GW markers is ~0.72 in the IBS kinship matrix, while the rest of the correlation coefficients are above 0.9. When the analysis is restricted to EA samples only, the two clusters are formed based on whether or not the kinship coefficient represents a pair of individuals from the same group. In the right panel, the within-group refers to kinship coefficients computed using individuals from the same group (Fin-Fin, CEU-CEU, etc.) and the between-group refers to kinship coefficients computed using individuals from different groups (Fin-CEU, Fin-GBR etc.).

In both the EA + AA and EA only data set, the genomic control factor λ_{GC} of the unadjusted model

TABLE 1 1000 Genome individuals used for genotype simulation

EA population	N	AA population	N
FIN	99	ASW	61
CEU	95	ACB	96
GBR	90	LWK	97
IBS	107	ESN	99
TSI	107	GWD	113
Total	498	MSL	84
		YRI	107
		Total	657

Abbreviations: AA, African ancestry; ACB, African Caribbean in Barbados; ASW, Americans of African Ancestry in Southwest USA; CEU, Utah Residents with Northern and Western European Ancestry; EA, European ancestry; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GWD, Gambian in Western Divisions in the Gambia; IBS, Iberian Population in Spain; MSL, Mende in Sierra Leone; LWK, Luhya in Webuye, Kenya; TSI, Toscani in Italy; YRI, Yoruba in Ibadan, Nigeria.
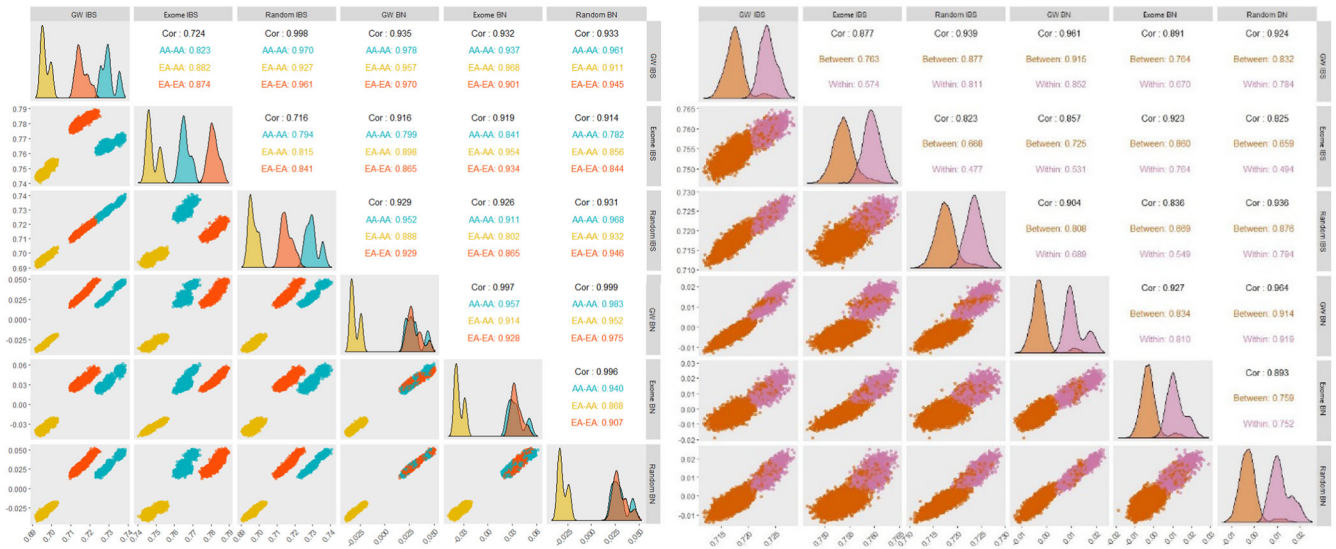
FIGURE 2 Pair-wise comparison of kinship coefficients computed using genome-wide and exonic SNVs. Left: kinship coefficients in the combined EA and AA samples. Right: kinship coefficients in EA samples only. Plots below the diagonal are the scatterplots of the kinship coefficients. Plots above the diagonal are the Pearson correlations between them. GW IBS and GW BN represent the IBS and BN kinship matrix with GW markers, Exome IBS and Exome BN are the IBS and BN kinship matrix with exonic markers, Random IBS and Random BN indicate the IBS and BN kinship matrix computed using a randomly selected subset of GW markers that has the same number of markers as the exonic set. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GW, genome-wide; IBS, identity-by-state; SNVs, single nucleotide variants

$Y \sim SNV$ indicates that there is population stratification present in the data, while in PC-adjusted models and LME/GLME models, λ_{GC} falls within an acceptable range except in GLME models using IBS kinship matrix alone: $Y \sim SNV + Kin_{GW,IBS}$ and $Y \sim SNV + Kin_{Exome,IBS}$ with binary outcome (Table 2). This indicates that PCs and BN GRM computed using either GW or exonic variants can correct the population stratification in both the binary and continuous data, whereas there is a slightly inflated λ_{GC} when using IBS GRM alone on a binary trait.

To examine Type-I error, we use 439,601 SNVs (EA + AA) and 516,250 SNVs (EA only) in each iteration

TABLE 2 Simulation result of median λ_{GC}

Model	Binary		Continuous	
	EA + AA	**EA only**	**EA + AA**	**EA only**
$Y \sim SNV$	43.98	6.117	16.15	2.783
$Y \sim SNV + Group$	1.003	1.003	1.000	0.999
$Y \sim SNV + PCs_{GW}$	1.006	1.007	0.999	0.999
$Y \sim SNV + PCs_{Exome}$	1.007	1.007	1.000	0.999
$Y \sim SNV + Kin_{GW, IBS}$	1.087	1.105	1.009	1.018
$Y \sim SNV + Kin_{Exome, IBS}$	1.063	1.087	1.003	1.005
$Y \sim SNV + Kin_{GW, BN}$	1.012	1.010	1.004	1.004
$Y \sim SNV + Kin_{Exome, BN}$	1.000	1.007	1.008	1.013
$Y \sim SNV + PCs_{GW} + Kin_{GW, IBS}$	1.004	1.001	–	–
$Y \sim SNV + PCs_{Exome} + Kin_{Exome, IBS}$	1.003	1.001	–	–
$Y \sim SNV + PCs_{GW} + Kin_{GW, BN}$	1.001	1.001	–	–
$Y \sim SNV + PCs_{Exome} + Kin_{Exome, BN}$	1.002	1.000	–	–

Abbreviations: AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GC, genomic control; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants.

and repeat 500 times for a total of 219,800,500 (EA + AA) and 258,125,000 (EA only) unassociated SNVs. Type-I error rate is computed under four different significance levels: 0.05, 1×10^{-3}, 1×10^{-4}, and 1×10^{-6} (Figures 3 and 4). For the binary outcome, there is a deflation in Type-I error rate in all models except the unadjusted model and GLME models using the IBS kinship matrix alone. However, the rest of the models have similar Type-I error rate as the gold standard model, which shows that they can also correctly control the Type-I error. For the continuous outcome, the PC-adjusted models correctly control Type-I error. In LME models, GRM computed using exonic markers has higher Type-I error rate than GRM computed using GW markers. While the relative Type-I error rates of both the GW- and exome-computed GRMs fall in an acceptable range at significance level 0.05, 1×10^{-3}, and 1×10^{-4}, there is a small inflation in exome-computed IBS GRM when the threshold is 1×10^{-6}. The extent of inflation in the EA + AA analysis is less than that in the EA only analysis, which indicates that the IBS GRM may not be sufficient to correct Type-I error when the population stratification is more subtle. In either the binary or the continuous outcome, models including PCs computed using GW or exonic variants do not show different Type-I error rate.

Power is evaluated using 30 SNVs that have different levels of confounding by population stratification and are associated with the simulated continuous or binary phenotypes (Figures 5 and 6). In GLME models for binary outcome, we focus on using the BN kinship matrix due to the inflated Type-I error rate found in models with the IBS kinship matrix. For the same reason, we do not include the unadjusted model in our power evaluation. We do not include the GLME models with PC adjustment in Figure 5 because the results are very similar to results from models using PC adjustment alone (Figure A1 in Appendix). We first compare power using GW-computed PCs/GRM versus exome-computed PCs/GRM. The empirical power evaluations are very similar using GW-computed PCs/GRM and exome-computed PCs/GRM. Then we evaluate the performance of the PC-adjusted and LME/GLME models when testing SNVs that have different levels of confounding by population stratification. In PC-adjusted models (binary or quantitative outcomes) and GLME models with PC adjustment (binary outcome, Figure A1 in Appendix), power is higher to detect low and medium weight SNVs (0th, 25th, 50th, and 75th percentile of PC weights, not or medium confounded by population stratification) than high weight SNVs (90th and 100th percentile of PC weights, highly confounded by population stratification) in general. This is expected because high weight SNVs contribute more to PCs than low and medium weight SNVs and hence the PCs can explain some of the phenotypic
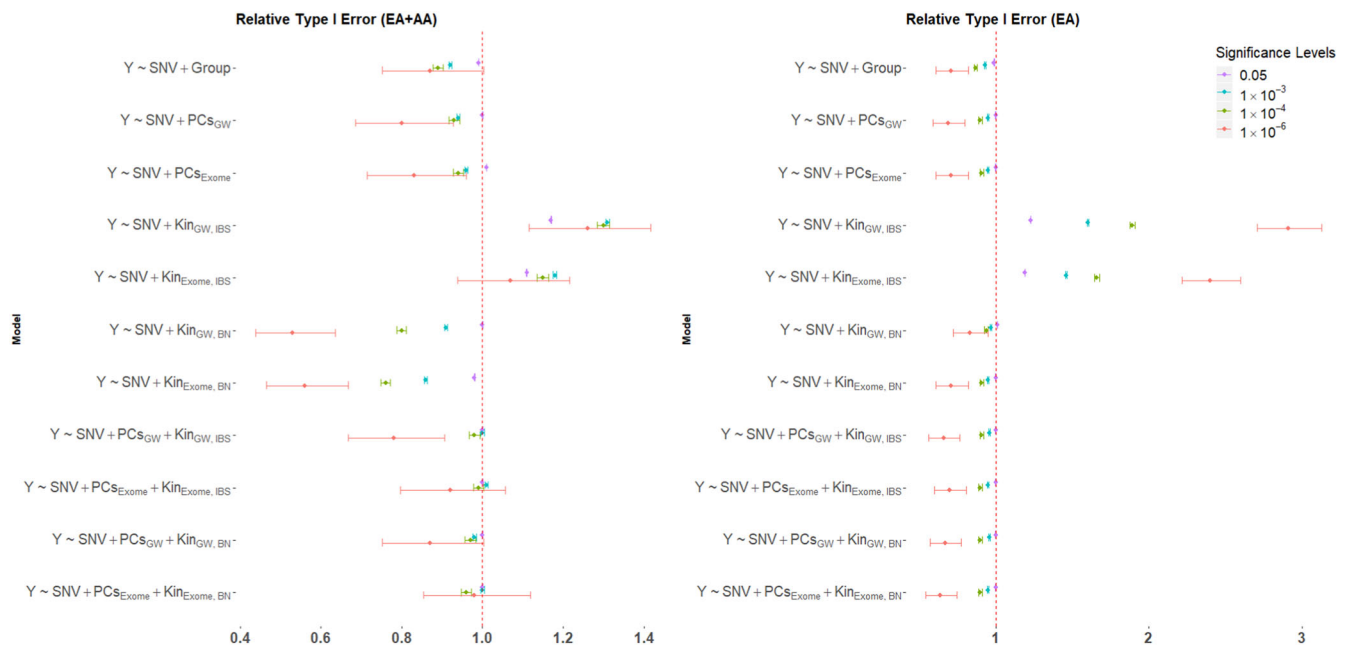

FIGURE 3 Relative Type-I error rate in simulation studies with binary outcome. A ratio of observed Type-I error rate to expected Type-I error rate for various *p*-value thresholds are presented. A ratio > 1 shows inflation and a ratio < 1 shows deflation. The unadjusted model Y ~ SNV has relative Type-I error rate 16, 645, 5724, and 449000 in the EA + AA analysis and 9, 179, 1101, and 42400 in the EA only analysis at $\alpha = 0.05$, 1×10^{-3}, 1×10^{-4}, and 1×10^{-6}, respectively. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants

FIGURE 4 Relative Type-I error rate in simulation studies with continuous outcome. A ratio of observed Type-I error rate to expected Type-I error rate for various p-value thresholds are presented. A ratio > 1 shows inflation and a ratio < 1 shows deflation. The unadjusted model Y ~ SNV has relative Type-I error rate 13, 383, 2840, and 147000 in the EA + AA analysis and 5, 48, 190, and 3026 in the EA only analysis at $\alpha = 0.05$, 1×10^{-3}, 1×10^{-4}, and 1×10^{-6}, respectively. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants
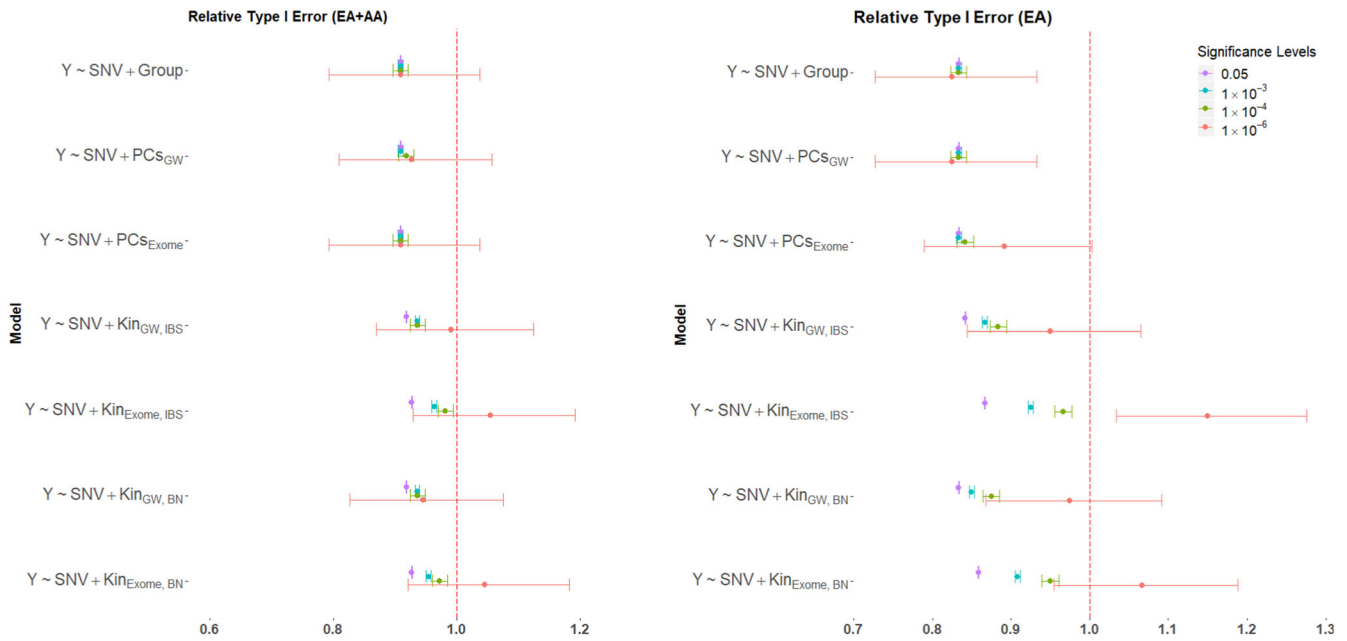

FIGURE 5 Power result from simulations with binary outcome when $\alpha = 1 \times 10^{-3}$. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants
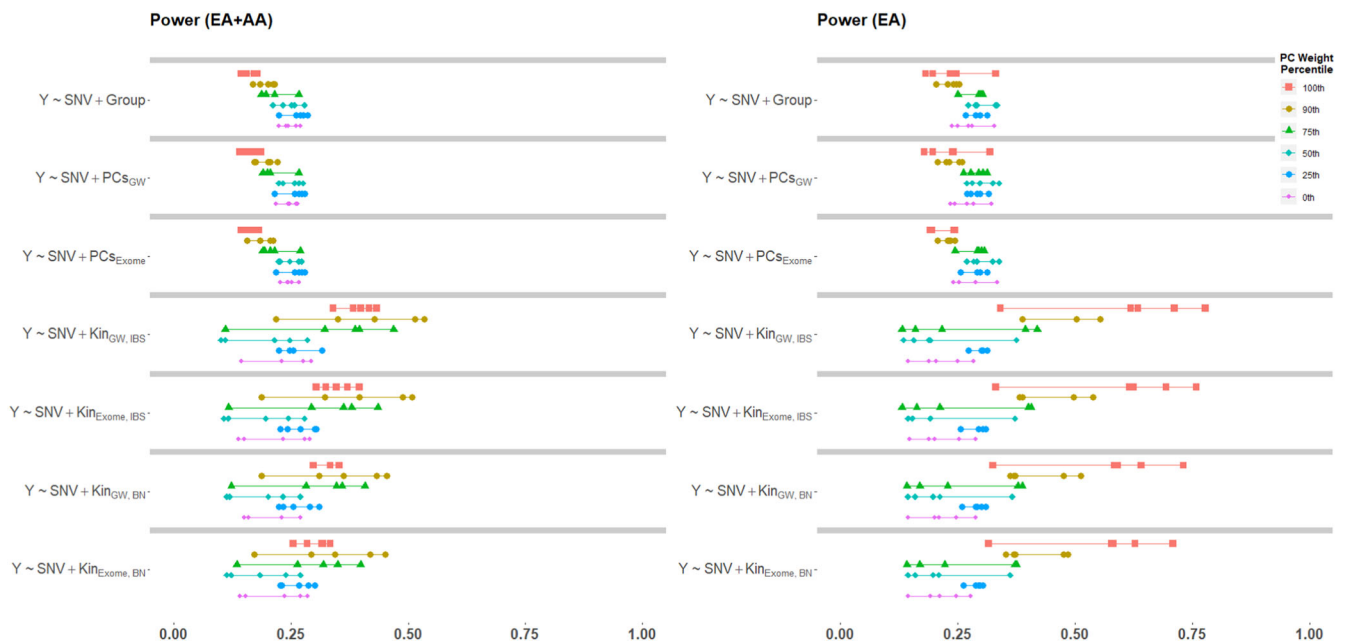
variance when testing high weight SNVs, which in turn decreases power in association tests. In mixed-effects models, high weight SNVs achieve similar or higher power than low and medium weight SNVs. This pattern is consistent across simulations with binary and continuous outcomes, and of EA and AA samples, and EA samples only. Finally, we directly compare the performance of PC-based versus LME/GLME models. When considering high weight SNVs which are at 90th and 100th percentile of PC weights, LME and GLME models
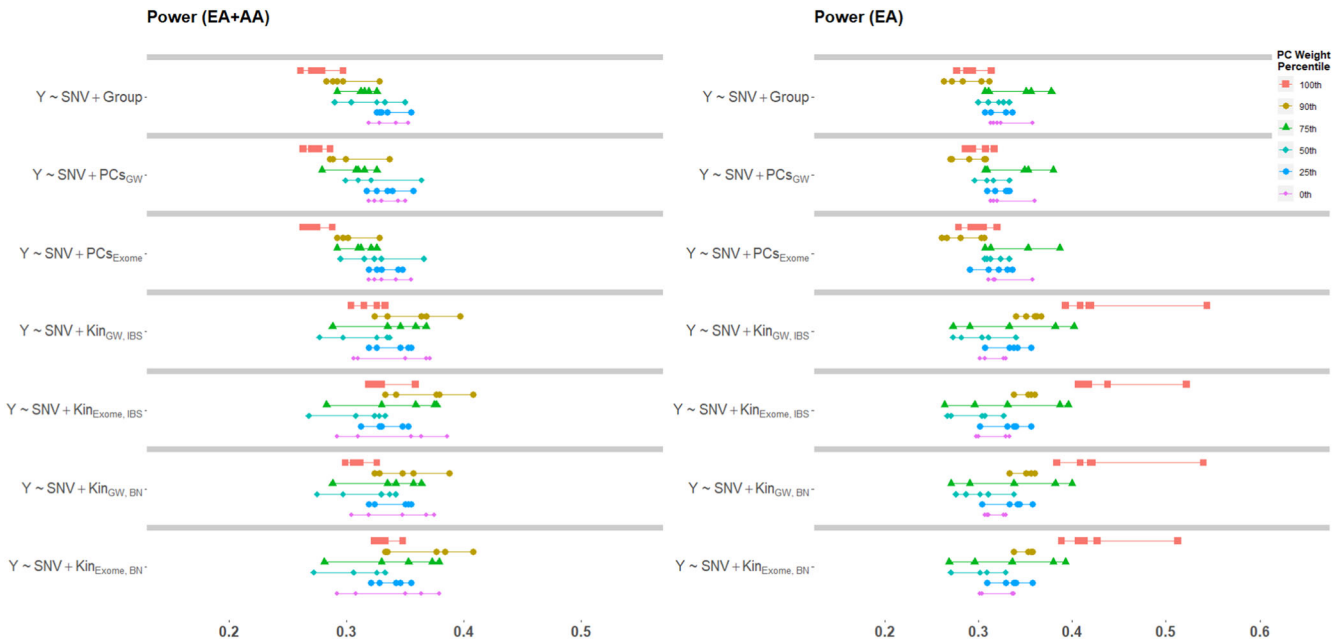
FIGURE 6 Power result from simulations with quantitative outcome when $\alpha = 1 \times 10^{-3}$. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants

using GRM alone outperform PC-based models in both the analysis with EA and AA individuals and the analysis restricted to EA individuals. For low and medium weight SNVs that are at 0th, 25th, 50th, and 75th percentile of PC weights, PC-based and LME models have similar power in general for both binary and continuous phenotypes. In addition, GLME models using both GRM and PC adjustment have similar power to the PC-based models for binary outcome.

In the application to FHS data, a subset of 122,233 GW SNVs and 18,107 exome chip SNVs pass the filtering criteria. A total of 2,464 unrelated individuals are selected to compute PC weights based on the known pedigree structures. The association analyses include 7,269 individuals of European ancestry. The genomic control factors are well-controlled except in the unadjusted model (Table 3). p-values for association between the *LCT* SNV rs2322659 and height in the unadjusted model

TABLE 3 Association analysis result in FHS real data example

Model	β	*SE*	*p*-value	λ_{GC}
Adjusted height ~ SNV	−.1343	0.017	1.6×10^{-15}	2.79
Adjusted height ~ SNV + $\text{Kin}_{\text{Pedigree}}$	.0893	0.042	.03	1.053
Adjusted height ~ SNV + $\text{Kin}_{\text{Pedigree}}$ + PCs_{GW}	.0031	0.046	.95	1.004
Adjusted height ~ SNV + $\text{Kin}_{\text{Pedigree}}$ + PCs_{EC}	−.0019	0.046	.97	1.004
Adjusted height ~ SNV + $\text{Kin}_{\text{GW,IBS}}$	.0220	0.020	.27	1.039
Adjusted height ~ SNV + $\text{Kin}_{\text{EC,IBS}}$	.0248	0.019	.19	1.045
Adjusted height ~ SNV + $\text{Kin}_{\text{GW,BN}}$	.0190	0.020	.34	1.020
Adjusted height ~ SNV + $\text{Kin}_{\text{EC,BN}}$	.0186	0.019	.33	1.045
Adjusted height ~ SNV + $\text{PC-AiR}_{\text{GW}}$ + $\text{Kin}_{\text{PC-Relate}}$	−.0108	0.020	.60	1.009
Adjusted height ~ SNV + $\text{PC-AiR}_{\text{EC}}$ + $\text{Kin}_{\text{PC-Relate}}$	−0.0061	0.019	.75	1.013

Abbreviations: EC, exome chip; GC, genomic control; GW, genome-wide; FHS, Framingham Heart Study; IBS, identity-by-state; PCs, principal components; *SE*, standard error; SNV, single nucleotide variant.

and LME model with a kinship matrix computed using the known family structure are 1.6×10^{-15} and 0.03, respectively, which indicates SNV rs2322659 is associated with height when not adjusting for population structure. However, this association is no longer statistically significant in PC-adjusted models, LME models using an empirical kinship matrix, and LME models using PCs and GRM computed by PC-AiR and PC-Relate (Table 3). This confirms that the spurious association in the unadjusted model is due to population stratification. Kinship matrix based on known pedigree structures is not sufficient to capture population stratification, whereas empirical kinship matrices based on genotypes are able to correct for both the family structure and population stratification. *p*-values of the PC-adjusted models and LME models using an empirical kinship matrix are all above 0.05, which shows that the adjustment using exome chip variants can also alleviate the spurious association due to population stratification in practice.

4 | DISCUSSION AND CONCLUSION

We compare PCs and GRM computed using GW and exonic SNVs, and evaluate model performance of PC-based and mixed-effects-model-based methods in terms of Type-I error rate and power in association tests of common variants with additive effect. Intuitively, in studies with exonic data, adjustment for potential population stratification may not be achieved because exonic SNVs only contain ancestry information within the exome. Exome-computed PCs and GRM may not be able to capture all information available in whole genome data. In addition, the smaller number of exonic variants may lead to insufficient adjustment of population stratification because the best ancestry estimates are usually obtained using a very large number of random markers (Price et al., 2010). However, our simulation and real data example showed that exonic markers can achieve a similar performance as GW markers for population stratification adjustment.

Through the comparison between PCs computed from GW and exonic markers using the Wilcoxon signed-rank test, we found that the significant difference was due to the difference in the number of GW and exonic SNVs included in PCA. When we used the same number of GW and exonic variants to compute PCs, the difference diminished. Comparison among kinship demonstrated that the BN GRM showed consistent high correlation (~0.99) between GW and exonic markers, while GW- and exome-computed IBS GRMs were less correlated (~0.72). This suggests that the exome-computed IBS GRM may

not perform as well as the GW-computed IBS GRM in association testing, which was verified in our Type-I error simulations.

Genomic control factor is often used to examine the inflation at the median of the null distribution, while Type-I error rate is used to examine the tail of the null distribution. We compared both quantities through simulation and drew different conclusions about the ancestry adjustment using exonic variants. The genomic control factor showed no evidence of inflation in models using either GW- or exome-computed PCs or GRM (except IBS GRM in binary trait). However, a slightly inflated Type-I error rate was found in LME models (continuous trait) with exome-computed IBS GRM at the $\alpha = 1 \times 10^{-6}$ significance level in the EA only analysis. In contrast to the exome-computed IBS GRM, the exome-computed BN GRM and the GW-computed IBS/BN GRM showed no inflation in the Type-I error rate and had similar performance as the PC-adjusted models in terms of λ_{GC} and Type-I error. These results reflect the moderate correlation ~0.72 between GW- and exome-computed IBS kinship coefficients. Due to the inflated λ_{GC} and Type-I error rate in models with IBS GRM, we recommend the use of BN GRM in LME/GLME models.

Our power evaluation was conducted using 30 SNVs that have different levels of confounding by population stratification. The power of using exome-computed PCs/GRM was very close to GW-computed PCs/GRM, which indicates that it is appropriate to use exonic data to detect SNVs with true effect. We acknowledge that our assumption of $R^2 = 0.5\%$ is large. This assumption was made due to computational constraints which necessitated a modest sample size for simulations. However, we could achieve the same power with a lower R^2 but larger sample size.

PC-adjusted models had higher power for SNVs not confounded by population stratification and lower power for SNVs highly confounded by population stratification. High weight SNVs contributed more to PCs than low and medium weight SNVs and hence PCs can explain part of the phenotypic variance when testing high weight SNVs, which in turn decreases the power in association analyses. For both continuous and binary outcomes, mixed-effects models had similar power as PC-based models in tests of low and medium weight SNVs, and higher power when testing high weight SNVs. Hence, they are more appropriate in association analyses on high weight SNVs due to their better performance.

Although using GW variants to compute PCs should be the preferred approach because this approach captures variation over the whole genome, exome-computed PCs are sufficient to control inflated Type-I error due to population stratification and provide similar power to

approaches adjusting for GW-computed PCs. LME/GLME models should be preferred over PC-adjusted models if associated SNVs are highly confounded with population stratification, such as SNVs in the HLA regions, which are associated with many autoimmune traits but also show signs of population stratification.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available in the database of Genotypes and Phenotypes (phs000007.v30.p11).

ORCID
Yuning Chen http://orcid.org/0000-0002-7358-7055
Gina M. Peloso https://orcid.org/0000-0002-5355-8636
Ching-Ti Liu http://orcid.org/0000-0002-0703-0742
Anita L. DeStefano https://orcid.org/0000-0002-9335-2819
Josée Dupuis https://orcid.org/0000-0003-2871-3603

REFERENCES
Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., ... Abecasis, G. R. 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Belkadi, A., Pedergnana, V., Cobat, A., Itan, Y., Vincent, Q. B., Abhyankar, A., ... Laurent, A. (2016). Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(24), 6713–6718. https://doi.org/10.1073/pnas.1606460113

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Chen, W., Chen, H., Wang, C., Conomos, M., Stilp, A., Li, Z., ... Lin, X. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, *98*(4), 653–666. https://doi.org/10.1016/j.ajhg.2016.02.012

Collins, A. R. (2007). *Linkage Disequilibrium and Association Mapping* (pp. 1–255). Totowa, NJ: Humana Press.

Conomos, M. P., Gogarten, S. M., Brown, L., Chen, H., Rice, K., Sofer, T., ... Yu, C. (2019). *GENESIS: GENetic EStimation and inference in structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness*. Retrieved from https://github.com/UW-GAC/GENESIS

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, *98*(1), 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, *39*(4), 276–293. https://doi.org/10.1002/gepi.21896

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Eu-Ahsunthornwattana, J., Miller, E. N., Fakiola, M., Jeronimo, S. M. B., Blackwell, J. M., & Cordell, H. J. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics*, *10*(7), e1004445. https://doi.org/10.1371/journal.pgen.1004445

Gazal, S., Gosset, S., Verdura, E., Bergametti, F., Guey, S., Babron, M., & Tournier-Lasserve, E. (2016). Can whole-exome sequencing data be used for linkage analysis? *European Journal of Human Genetics*, *24*(4), 581–586. https://doi.org/10.1038/ejhg.2015.143

Groop, L. C., Campbell, C. D., Lunetta, K. L., Lyon, H. N., Altshuler, D., Hirschhorn, J. N., ... Ogburn, E. L. (2005). Demonstrating stratification in a european american population. *Nature Genetics*, *37*(8), 868–872. https://doi.org/10.1038/ng1607

Grove, M., Yu, B., Cochran, B., Haritunians, T., Bis, J., Taylor, K., ... Boerwinkle, E. (2013). Best practices and joint calling of the HumanExome BeadChip: The CHARGE consortium. *PLoS One*, *8*(7), e68095. https://doi.org/10.1371/journal.pone.0068095

Hirschhorn, J. N., Pato, C. N., Henderson, B., Petryshen, T. L., Altshuler, D., Reich, D., ... Gabriel, S. B. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, *36*(4), 388–393. https://doi.org/10.1038/ng1333

Holland, J. B., Buckler, E. S., Briggs, W. H., Gaut, B. S., Kresovich, S., McMullen, M. D., ... Vroh Bi, I. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

International HapMap Consortium. (2003). The international HapMap project. *Nature*, *426*(6968), 789–796. https://doi.org/10.1038/nature02168

Kancheva, D., Atkinson, D., De Rijk, P., Zimon, M., Chamova, T., Mitev, V., ... Jordanova, A. (2016). Novel mutations in genes causing hereditary spastic paraplegia and charcot-marie-tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. *Genetics in Medicine*, *18*(6), 600–607. https://doi.org/10.1038/gim.2015.139

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, *42*(4), 348–354. https://doi.org/10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, *165*(4), 2213–2233.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. https://doi.org/10.1038/nmeth.1681

Loh, P., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., ... Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*(3), 284–290. https://doi.org/10.1038/ng.3190

Ma, C., Blackwell, T., Boehnke, M., & Scott, L. J. GoT2D Investigators (2013). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, *37*(6), 539–550. https://doi.org/10.1002/gepi.21742

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Price, A. L., Zaitlen, N. A., Patterson, N., & Reich, D. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459–463. https://doi.org/10.1038/nrg2813

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. Retrieved from http://www.genetics.org/cgi/content/abstract/155/2/945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, *34*(28), 3769–3792. https://doi.org/10.1002/sim.6605

Smith, K. R., Bromhead, C. J., Hildebrand, M. S., Shearer, A. E., Lockhart, P. J., Najmabadi, H., ... Bahlo, M. (2011). Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biology*, *12*(9), R85. https://doi.org/10.1186/gb-2011-12-9-r85

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477. https://doi.org/10.1371/journal.pgen.1000477

Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics*, *27*(16), 2304–2305. https://doi.org/10.1093/bioinformatics/btr341

Tucker, G., Price, A. L., & Berger, B. (2014). Improving the power of GWAS and avoiding confounding from population stratification with PC-select. *Genetics*, *197*(3), 1045–1049. https://doi.org/10.1534/genetics.114.164285

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, K. (2009). Testing for genetic association in the presence of population stratification in genome-wide association studies. *Genetic Epidemiology*, *33*(7), 637–645. https://doi.org/10.1002/gepi.20415

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Zhang, Z., Gore, M. A., Yu, J., Buckler, E. S., Tiwari, H. K., Ordovas, J. M., ... Lai, C. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., ... Nordborg, M. (2007). An arabidopsis example of association mapping in structured samples. *PLoS Genetics*, *3*(1), e4. https://doi.org/10.1371/journal.pgen.0030004

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310
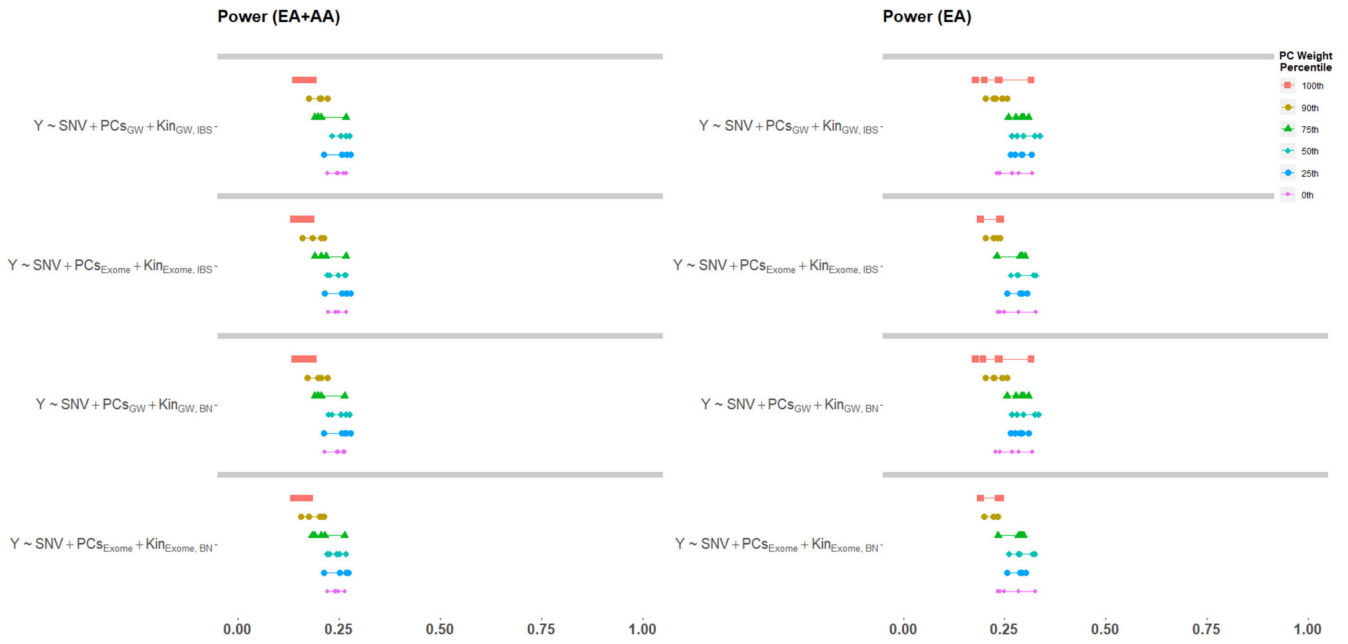
APPENDIX

See Figure A1 and Table A1

FIGURE A1 Power result of GLME models with PC adjustments from simulations with binary outcome when $\alpha = 1 \times 10^{-3}$. AA, African ancestry; BN, Balding–Nichols; EA, European ancestry; GLME, generalized linear mixed-effects; GW, genome-wide; IBS, identity-by-state; PCs, principal components; SNVs, single nucleotide variants

TABLE A1 List of SNVs used in simulation studies of power

PC weight percentile	EA + AA		EA only	
	RS ID	CHR:POS:REF:ALT	RS ID	CHR:POS:REF:ALT
0th	rs11084445	19:56823708:C:A	rs1047406	8:22570935:T:C
	rs6480463	10:72517837:C:T	rs2276232	21:43291997:C:A
	rs1130169	4:15850685:C:T	rs3745009	18:43262359:G:A
	rs3019086	8:101393896:A:G	rs161557	5:143200053:C:T
	rs3814018	1:95086778:C:T	rs3746619	20:54823805:C:A
25th	rs1041316	14:57099859:G:A	rs1061409	3:145917761:T:C
	rs10815567	9:733049:G:A	rs12442757	15:65689267:T:C
	rs4531246	1:2362020:G:T	rs2072409	7:31746837:C:T
	rs3814055	3:119500035:C:T	rs12460981	19:10628548:A:G
	rs6574791	14: 20731130:C:T	rs677595	18:65541451:G:A
50th	rs7909074	10:45395839:G:A	rs2273660	2:32961858:A:G
	rs8239	12:118502467:A:G	rs11151964	18:72201918:G:A
	rs409782	21:15596772:T:G	rs2240795	5:149589647:A:G
	rs1805034	18:60027241:C:T	rs505187	19:56588609:G:A
	rs2279651	5:35039437:A:G	rs12051	5:35039437:A:G
75th	rs6573560	14:65031534:C:T	rs4745571	9:79318367:T:C
	rs3826537	17:60769803:A:G	rs1581688	7:112723093:T:C
	rs3742764	14:76330181:G:A	rs10505093	8:107617360:A:G
	rs464333	21:41099821:G:A	rs2243193	1:207016225:A:G
	rs2452524	15:50226313:G:T	rs2279303	8:49536412:G:T
90th	rs1052690	9:86258685:A:C	rs13874	3:66419956:C:T
	rs1407977	9:15188106:T:C	rs3750306	8:139164192:G:A
	rs9911122	17:74208424:G:A	rs2074412	17:35839066:C:T
	rs289723	16:57080528:C:A	rs4129081	2:191307408:G:T
	rs2239923	17:43176804:C:T	rs1133099	16:2015121:C:T
100th	rs3820416	1:170707675:T:C	rs3783501	19:2477316:G:A
	rs1567047	4:126372742:G:A	rs3741190	11:66816507:T:C
	rs11155297	6:143825104:G:T	rs9559516	13: 109859945:T:C
	rs10933164	2:227860671:G:A	rs2071593	6:31512799:G:A
	rs3741194	11:66626234:T:C	rs764231	15:34943414:G:A

Note: Positions are reported on genome build hg19.

Abbreviations: AA, African ancestry; EA European ancestry; SNVs, single nucleotide variants.